

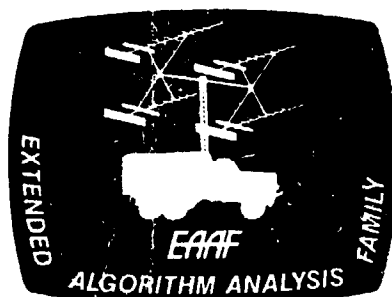
7057-015, Rev. A

U.S. ARMY INTELLIGENCE CENTER AND SCHOOL  
SOFTWARE ANALYSIS AND MANAGEMENT SYSTEM

## FUNDAMENTALS OF LINEAR ESTIMATION

### TECHNICAL MEMORANDUM No. 17

Charles L. Lawson



28 September 1987

National Aeronautics and  
Space Administration

**JPL**

JET PROPULSION LABORATORY  
California Institute of Technology  
Pasadena, California

JPL D-4305, Rev. A  
ALGO\_PUB\_0031

DTIC  
ELECTE  
JUN 23 1988  
S E D

This document has been approved  
for public release and sales in  
distribution is unlimited.

88 6 22 007

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ALGO PUB 0031	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle)  Technical Memo 17, rev A, "Fundamentals of Linear Estimation"		5. TYPE OF REPORT & PERIOD COVERED  FINAL
		6. PERFORMING ORG. REPORT NUMBER D-4305, rev A
7. AUTHOR(s) Dr. Charles L. Lawson		8. CONTRACT OR GRANT NUMBER(s)  NAS7-918
9. PERFORMING ORGANIZATION NAME AND ADDRESS Jet Propulsion Laboratory, ATTN: 171-209 California Institute of Technology 4800 Oak Grove, Pasadena, CA 91109		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS  RE 182 AMEND #187
11. CONTROLLING OFFICE NAME AND ADDRESS Commander, USAICS ATTN: ATSI-CD-SF Ft. Huachuca, AZ 85613-7000		12. REPORT DATE 28 Sep 87
		13. NUMBER OF PAGES 39
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)  Commander, USAICS ATTN: ATSI-CD-SF Ft. Huachuca, AZ 85613-7000		15. SECURITY CLASS. (of this report)  UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE NONE
16. DISTRIBUTION STATEMENT (of this Report)  Approved for Public Dissemination		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES  Prepared by Jet Propulsion Laboratory for the US Army Intelligence Center and School's Combat Developer's Support Facility.		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number)  Fix Estimation, Linear Operator, Mean, Variance, One Dimensional Estimation, Multi-Dimensional Estimation, Matrices, Cholesky Decomposition, QR Decomposition, Distributions, Least Squares Estimation		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number)  This report presents the fundamentals of linear estimation theory as needed for fixing. The initial chapters are devoted to the one dimensional case. After introducing concepts from linear algebra, the multidimensional case is discussed. Material from statistical distributions is introduced as needed. The final chapter is con- cerned with combining dissimilar data sets to obtain new estimates. Two minor errors were noted. On page 4, the last 2 in the sixth paragraph should be .02. $A^*$ , the <u>transpose</u> of $A$ , is not defined.		

U.S. ARMY INTELLIGENCE CENTER AND SCHOOL  
Software Analysis and Management System

FUNDAMENTALS OF LINEAR ESTIMATION

EAAF

Technical Memorandum No. 17

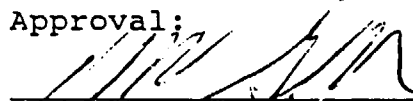
28 September 1987

Author:



Charles L. Lawson, Supervisor  
Applied Mathematics Group

Approval:




James W. Gillis, Subgroup Leader  
Algorithm Analysis Subgroup



Edward J. Records, Supervisor  
USAMS Task

Concur:



A. F. Eilman, Manager  
Ground Data Systems Section



Fred Vote, Manager  
Advanced Tactical Systems

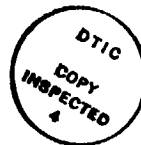
JET PROPULSION LABORATORY  
California Institute of Technology  
Pasadena, California

## PREFACE

The work described in this publication was performed by the Jet Propulsion Laboratory, an operating division of the California Institute of Technology, under contract NAS7-918, RE182, A187 with the National Aeronautics and Space Administration, for the United States Army Intelligence Center and School.

This revised edition replaces the original report, published on June 26, 1987, and is being re-published because the original edition contains photocopy blotches and illegible print.

Accession For	
NTIS GRA&I	<input checked="" type="checkbox"/>
DTIC TAB	<input checked="" type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	



## EXECUTIVE SUMMARY

This Technical Memorandum was prepared originally as part of the Generic Fix Report (FY-86) which was eliminated under the FY-87 statement of work (SOW #2), undated (delivered to JPL 19 November 1986).

The purpose of the Generic Fix Report, of which this paper was to be an appendix, was to collect all the material needed to understand Direction Finding and Fix Estimation and their mathematical basis in one volume to support the multi-volume series of Fix Estimation Reports.

This paper is being published because it was completed in FY-86 with FY-86 funds and was being held for integration into the Generic Fix Report. It will be of value to readers desiring to pursue the mathematics involved in the Fix Estimation Reports.

FUNDAMENTALS OF LINEAR ESTIMATION

Charles L. Lawson

Section 366

Computing Memorandum No. 518

August 1, 1986

Jet Propulsion Laboratory  
California Institute of Technology  
Pasadena, California

## TABLE OF CONTENTS

IVE SUMMARY	
roduction . . . . .	1
dimensional estimation . . . . .	2
1. A single observation . . . . .	2
2. Two observations . . . . .	2
3. Two observations with differing precision . . . . .	3
acterizing random distributions . . . . .	3
mples of the use of the mean and standard deviation . . . . .	6
1. Example assuming $\sigma$ is known . . . . .	6
2. Example assuming $\sigma$ is unknown . . . . .	8
multidimensional linear estimation problem . . . . .	9
1. Notation and concepts of linear algebra . . . . .	9
2 N-dimensional random variables . . . . .	12
3. Linear estimation of n parameters using m observations . . . . .	13
4.3.1. The estimator $\hat{\xi}$ and its covariance matrix . . . . .	13
4.3.2. The estimator $\hat{\eta}$ and its covariance matrix . . . . .	15
4.3.3. Interpretation as least squares estimation . . . . .	15
4.3.3.1. Transformations of the least squares problem . . . . .	16
4.3.4. Introducing a scaling factor for H . . . . .	18
4.3.4.1. An unbiased estimator for $\phi$ . . . . .	19
andard distributions . . . . .	20
5.1. The normal or Gaussian distribution . . . . .	20
5.2. Confidence regions . . . . .	21
5.2.1. Geometric characterization of a confidence ellipsoid . . . . .	22
5.3. The $\chi^2$ distribution . . . . .	22
5.4. Student's t distribution . . . . .	24
5.5. The F distribution . . . . .	26
approaches to data analysis . . . . .	26
6.1. Analysis of one coherent set of data . . . . .	26
6.2. Combining sets of data . . . . .	28
6.2.1 Addition. Remarks on combining data sets . . . . .	30
ENCES . . . . .	32
INDEX . . . . .	33

## Fundamentals of Linear Estimation

### 0. Introduction

This memorandum will treat a selection of topics in linear estimation, beginning with a very simple situation and progressing through more complications. These topics are shown in A. below. This progression will provide the main structure for the document. Topics from B. and C. below will be brought in at points where their need has been motivated.

A more conventional academic approach to covering this material would be to discuss B. and C. first, to establish the foundation, and then treat A. This approach would be easier to do but would lack motivation in the early stages. We are trying the stated approach on the assumption that the sponsor wants a more motivated presentation.

#### A. Topics in estimation;

- 1) One dimensional estimation with one observation,
- 2) Same with two observations,
- 3) With same or different observational errors,
- 4) Any number of observations,
- 5) Two dimensional estimation, independent errors,
- 6) Multidimensional estimation,
- 7) Combining sets of observations;

#### B) Properties of random distributions!

- 1) Mean, first moment,
- 2) Standard deviation, variance, second moment,
- 3) Distribution, frequency function, all moments, -  
Normal distribution,  
Chi squared distribution,  
Student's t distribution,  
F distribution.
- 4) Confidence intervals;

#### C) Least squares;

- 1) Statement, geometric interpretation, -  
Gradient of the sum of squares,
- 2) Solution methods, -  
Orthogonal transformations, and
- 3) Using the covariance matrix of observation errors. (Lagrange: 5th edition)

The general style of the paper is tutorial, however due to the amount of material being covered, and to avoid reaching book-length, it will be more of a sketch of a tutorial rather than a true tutorial in some places.



Equations will be numbered within sections. For example, if there is an Equation (2) in Section 4.1, it will be referenced as Eq.(2) from within Section 4.1 and as Eq. 4.1.(2) from any other section.

## 1. One dimensional estimation

### 1.1. A single observation

Suppose one makes a one dimensional observation, such as the distance between two stakes at a construction site. Suppose the measured distance is 95.36 meters, with an uncertainty of 2 centimeters.

For the moment we shall not define this concept of uncertainty too precisely. One way to think of measurement uncertainty is in terms of how surprised we would be at different possible outcomes if we could somehow later determine the distance much more accurately. We would be surprised if our error turned out to be 3 cm., and very surprised if it was 4 cm., and not at all surprised if it was just 1 cm. If the error turned out to be greater than about 6 cm. we would probably check to see if there was a blunder in the first measurement or if the stakes had moved.

### 1.2. Two observations

Suppose we make a second observation of this same distance and obtain 95.37 meters, again with an uncertainty of 2 cm. What is our best estimate of the true distance?

If we use the principle of least squares, which we will not justify at this point, we seek a number,  $x$ , such that the sum of squares of residuals between  $x$  and the observed values is minimized. Thus denoting the measurements by  $b_1 = 95.36$  and  $b_2 = 95.37$ , we seek  $x$  to minimize

$$s = (x - b_1)^2 + (x - b_2)^2$$

We may differentiate  $s$  with respect to  $x$ , obtaining

$$ds/dx = 2(x - b_1) + 2(x - b_2)$$

which will have the value zero when

$$x = (b_1 + b_2)/2$$

i.e., when  $x$  is the average or mean of  $b_1$  and  $b_2$ . Thus our estimate of the distance being measured is 95.365 meters.

What estimate of uncertainty do we attach to this result? To answer this we shall need to adopt a mathematical model of uncertainty, but before doing this we shall introduce one more example.

### 1.3. Two observations with differing precision

Suppose we have the first observation as before but then make a second observation of this same distance using a more precise measuring device, obtaining a distance of 95.372 meters, with an uncertainty of 0.2 cm. Now the simple average of the two measurements no longer seems like a reasonable estimate. We should somehow give more weight to the more accurate measurement.

A reasonable way to do this is to scale the residuals for the two measurements so that equal values of the scaled residuals correspond to equal levels of surprise. Specifically, instead of the simple residuals,  $(x - b_1)$  and  $(x - b_2)$ , we will use the scaled residuals,  $(x - b_1)/d_1$  and  $(x - b_2)/d_2$ , where  $d_1$  denotes the uncertainty in the measurement  $b_1$ . Thus, for example, if  $(x - b_1)/d_1$  has the value 1.2, this engenders the same level of surprise as would be associated with  $(x - b_2)/d_2$  having the value 1.2.

The combined error function we shall now seek to minimize is

$$s = [(x - b_1)/d_1]^2 + [(x - b_2)/d_2]^2$$

Differentiating with respect to  $x$  we obtain

$$ds/dx = 2[(x - b_1)/d_1 + (x - b_2)/d_2]$$

which will have the value zero when

$$x = (b_1/d_1 + b_2/d_2) / (1/d_1 + 1/d_2)$$

Using the values,  $b_1 = 95.36$ ,  $d_1 = 0.02$ ,  $b_2 = 95.372$ , and  $d_2 = 0.002$ , we obtain the estimate,  $x = 95.3709$ . Note that with this estimate the simple residuals are

$$x - b_1 = 0.0109$$

and

$$x - b_2 = -0.00109$$

whereas the scaled residuals have equal magnitudes of

$$|(x - b_1)/d_1| = |(x - b_2)/d_2| = 0.545$$

Looking on to larger problems, we remark that although least squares estimation has a tendency to balance the magnitudes of scaled residuals, the actual data and dimensionality of a problem limits how closely this balance can be approached, and in general one can not hope for exact balancing as was attained in this example.

Now we must develop a mathematical model for uncertainties.

## 2. Characterizing random distributions

Consider again our first example in which we assumed the uncertainty of the measurement was 2 cm. Suppose we repeat this measurement 1000 times and count the number of times the difference from our

original measurement falls in selected ranges, such as  $(-\infty, -4 \text{ cm.})$ ,  $(-4, -2)$ ,  $(-2, 0)$ ,  $(0, 2)$ ,  $(2, 4)$ , and  $(4, \infty)$ . If we repeated this measurement another 1000 times we would expect some stability in the percentage of measurement differences falling in each one of our bins. For example we would expect the percentage of measurement differences falling in  $(0, 2 \text{ cm.})$  would hover around some fixed value, say 34%.

A mathematical model that is useful in analyzing this type of behavior is the assumption that there is a nonnegative continuous function,  $f$ , defined for all real numbers, and related to our experiment by the condition that the area under the graph of  $f$  between any two points  $a$  and  $b$  gives the value to which this type of repeated experimenting and counting converges. Thus, such a function,  $f$ , relating to our experiment would need to have area between 0 and 2 of 0.34.

The usual statistical terminology is to call a function,  $f$ , a frequency function or a probability density function if it is nonnegative and its integral from  $-\infty$  to  $+\infty$  exists and has the value 1. We will only be concerned with frequency functions that are continuous, or at most have jump discontinuities at a finite number of points.

The indefinite integral of a frequency function is called a distribution function. Thus from a frequency function,  $f$ , we obtain a distribution function,  $F$ , defined by

$$F(t) = \int_{-\infty}^t f(s) \, ds$$

A distribution function is defined for all real numbers, is continuous and monotone nondecreasing. It approaches the limiting value of 0 as its argument approaches  $-\infty$ , and 1 as its argument approaches  $+\infty$ .

The term, random variable, is commonly used to refer to a quantity, such as the measurement error in our example, that typically has a different unpredictable value each time it is observed, but yet exhibits some regularity with regard to the distribution of its values in a large number of observations. Note that we are not actually giving a definition of the term, random variable.

The closest we can come to giving a mathematical definition of the term, random variable, is to say that the statement, " $x$  is a random variable with probability density function  $f$ " means that certain stylized statements involving " $x$ " are to be taken as meaning something specific about " $f$ ". As an example of such a statement, note that "the probability that  $x$  exceeds 2 is 0.02", which may also be expressed as " $P(x > 2) = 0.02$ ", means "the integral of  $f$  from 2 to  $+\infty$  is 0.02".

It is often convenient to use the term, distribution, as a linguistic aid in associating the name of a random variable with the name of its frequency function. For example we may at some point let  $D$  denote a random distribution with frequency function,  $f$ , and later say that  $\hat{x}$  is a sample from  $D$ .

In practice one almost never has enough information to determine a frequency function from empirical data. Various assumptions are typically made to fill this gap.

For some purposes it suffices just to assume that there is some frequency function underlying the random aspects of a problem and compute estimates of certain attributes of the distribution, most commonly the mean, which is a measure of the central location of the distribution, and the standard deviation, which is a measure of the dispersion of the distribution.

When one wishes to go further and make statements involving probabilities, it becomes necessary to base the analysis on some specific frequency function. There are a number of frequency functions that have been thoroughly studied by statisticians, so in practice one usually picks one of these well known functions that has a plausible shape for the problem.

The mean of a distribution with frequency function,  $f$ , is defined by

$$\mu = \int_{-\infty}^{\infty} s f(s) ds$$

while the standard deviation,  $\sigma$ , is defined by

$$\sigma^2 = \int_{-\infty}^{\infty} (s - \mu)^2 f(s) ds$$

The squared quantity,  $\sigma^2$ , is called the variance of the distribution.

It is useful to have notations for these concepts for use with the "random variable" terminology. Thus the mean value of a random variable,  $x$ , is also called the expected value of  $x$ , denoted by  $E(x)$ . This notation is extended to apply to arbitrary functions of a random variable. Thus if  $g(x)$  is any function of a random variable,  $x$ , and the following integral exists, we may write

$$E(g(x)) = \int_{-\infty}^{\infty} g(s) f(s) ds$$

The expected value operator is a linear operator, in the sense that for arbitrary scalars,  $\alpha$  and  $\beta$ , and functions,  $g$  and  $h$ , for which the required integrals exist, we have

$$E(\alpha g(x) + \beta h(x)) = \alpha E(g(x)) + \beta E(h(x))$$

Using the expected value notation the definition of the standard deviation,  $\sigma$ , can be written as

$$\sigma^2 = E( (x - E(x))^2 )$$

It is useful to introduce another operator,  $\text{Var}(x)$ , to capture this last expression. Thus we define

$$\text{Var}(x) = E( (x - E(x))^2 )$$

The value of the Var operator is insensitive to an additive shift of the underlying distribution and varies with the square of a multiplicative factor. Thus

$$\text{Var}(\alpha x + \beta) = \text{Var}(\alpha x) = \alpha^2 \text{Var}(x)$$

### 3. Examples of the use of the mean and standard deviation

#### 3.1. Example assuming $\sigma$ is known

Returning to the example of Sec. 1.2., let us model the uncertainty in the measurement process by assuming the observed values  $b_1$  and  $b_2$  are independent random samples from some distribution with frequency function,  $f$ , having mean,  $\mu$ , and standard deviation,  $\sigma$ . We assume that  $f$  and  $\mu$  are not known, but we make the rather strong assumption that  $\sigma$  is known to have the value 2 cm. Our goal is to estimate  $\mu$  and obtain an estimate of the standard deviation of the estimated value.

The estimation function used in Sec. 1.2. was the simple average

$$g(b_1, b_2) = (b_1 + b_2)/2$$

It will be instructive to consider a slightly more general estimator function, namely

$$h(b_1, b_2) = \alpha b_1 + \beta b_2$$

and then show that the choice of  $\alpha = \beta = 1/2$  has certain desirable properties.

Regarding  $b_1$  and  $b_2$  as independent random samples from our assumed distribution, the function  $h$  defines a new random variable having a different distribution. What are the mean and standard deviation of this derived distribution? What we hope, if  $h$  is to be of reasonable use as an estimator, is that the mean of  $h$  is  $\mu$ , or has a known functional relationship to  $\mu$ , and the standard deviation of  $h$  is less than  $\sigma$ , so we are estimating the quantity of interest,  $\mu$ , and with dispersion less than that of a single observation.

We must generalize the definitions given previously for the operators  $E()$  and  $\text{Var}()$ , because  $h$  depends on two random variables.

We compute the mean of  $h$  as

$$\begin{aligned} E(h(b_1, b_2)) &= \iint (\alpha b_1 + \beta b_2) f(b_1) f(b_2) db_1 db_2 \\ &= \alpha \int b_1 f(b_1) db_1 \int f(b_2) db_2 \\ &\quad + \beta \int f(b_1) db_1 \int b_2 f(b_2) db_2 \\ &= \alpha \mu + \beta \mu = (\alpha + \beta) \mu \end{aligned}$$

and the variance of  $h$  as

$$\begin{aligned} \text{Var}(h(b_1, b_2)) &= E([h - E(h)]^2) \\ &= E([\alpha b_1 + \beta b_2 - (\alpha + \beta) \mu]^2) \\ &= E([\alpha(b_1 - \mu) + \beta(b_2 - \mu)]^2) \\ &= \alpha^2 E[(b_1 - \mu)^2] + \beta^2 E[(b_2 - \mu)^2] + 2\alpha\beta E[(b_1 - \mu)(b_2 - \mu)] \\ &= \alpha^2 \text{Var}(b_1) + \beta^2 \text{Var}(b_2) \\ &\quad + 2\alpha\beta \iint (b_1 - \mu)(b_2 - \mu) f(b_1) f(b_2) db_1 db_2 \\ &= (\alpha^2 + \beta^2) \sigma^2 \\ &\quad + 2\alpha\beta \int (b_1 - \mu) f(b_1) db_1 \int (b_2 - \mu) f(b_2) db_2 \\ &= (\alpha^2 + \beta^2) \sigma^2 + 0 = (\alpha^2 + \beta^2) \sigma^2 \end{aligned}$$

For the simple average estimator,  $g$ , where  $\alpha = \beta = 1/2$ , these formulas give a mean value of  $\mu$  and a standard deviation of  $\sigma/2^{1/2}$  or 1.4 cm.

What about other values of  $\alpha$  and  $\beta$ ? An estimator is called unbiased if its mean value is equal to the quantity we wish to estimate, in this case,  $\mu$ . To achieve this we see that we must have  $\alpha + \beta = 1$ .

An estimator is called minimum variance within its class if no other estimator in its class has smaller variance. The minimum value of the factor  $(\alpha^2 + \beta^2)$ , subject to  $\alpha + \beta = 1$ , is attained when  $\alpha = \beta = 1/2$ .

Any estimator of the form  $\alpha b_1 + \beta b_2$  is called a linear estimator. From the above we see that such an estimator is an unbiased linear estimator if  $\alpha + \beta = 1$ , and it is the minimum variance unbiased linear estimator if  $\alpha = \beta = 1/2$ .

Remark: The term,  $E[(b_1 - \mu)(b_2 - \mu)]$ , in the above equation for  $\text{Var}(h(b_1, b_2))$  is called the covariance of  $b_1$  and  $b_2$ , and is denoted by  $\text{Cov}(b_1, b_2)$ . This is a very special case of the covariance since the joint frequency function of  $b_1$  and  $b_2$ , here,  $f(b_1)f(b_2)$ , is the product of two functions, each depending on only one of the variables. In such a case the covariance is zero because it can be written as the product

of two integrals, each of which is zero. We postpone discussion of covariance in more general situations until matrix notation has been introduced.

### 3.2. Example assuming $\sigma$ is unknown

Commonly one does not know  $\sigma$  apriori, as was assumed in Sec. 3.1., but rather needs to estimate it from the data. Such an estimate can be obtained from the sum of squares of residuals.

Define the residuals for the two measurements by

$$r_i = b_i - g(b_1, b_2), \quad i = 1, 2$$

and define

$$S^2 = r_1^2 + r_2^2$$

Each residual is a function of the random variables,  $b_1$  and  $b_2$ , and thus so is  $S^2$ . Therefore  $S^2$  is a derived random variable having its own distribution. The mean value of  $S^2$  may be determined as follows:

$$\begin{aligned} E(S^2) &= E(r_1^2 + r_2^2) \\ &= E([b_1 - (b_1 + b_2)/2]^2 + [b_2 - (b_1 + b_2)/2]^2) \\ &= 1/2 E([b_1 - b_2]^2) \\ &= 1/2 E([(b_1 - \mu) - (b_2 - \mu)]^2) \\ &= 1/2 (\text{Var}(b_1) + \text{Var}(b_2) - 2 \text{Cov}(b_1, b_2)) \\ &= 1/2 (\sigma^2 + \sigma^2 + 0) = \sigma^2 \end{aligned}$$

Thus  $S^2$  is an unbiased estimator for  $\sigma^2$ . In our example we have

$$S^2 = (-0.5 \text{ cm})^2 + (0.5 \text{ cm})^2 = 0.5 \text{ cm}^2$$

from which we obtain  $(0.5)^{1/2} = 0.71 \text{ cm}$  as an estimate of  $\sigma$ .

In the more general case of  $m$  observations and  $n$  parameters being estimated the expected value of  $S^2$  is  $(m-n)\sigma^2$ . Thus  $S^2/(m-n)$  is an unbiased estimator for  $\sigma^2$ . In our example we obtained  $E(S^2) = \sigma^2$  because we have  $m = 2$  and  $n = 1$ . The difference,  $(m-n)$ , is called the number of degrees of freedom in the problem.

To estimate the dispersion of  $S^2$  requires more information or more assumptions. The usual approach is to assume the distribution from which the data values,  $b_i$ , arise is a normal distribution. Then the scaled derived random variable  $S^2/\sigma^2$  will have a  $\chi^2$  distribution with  $m-n$  degrees of freedom. The normal distribution and  $\chi^2$  distribution will be defined in Sec. 5.

Discussion along the above lines could be carried out for the case of Sec. 1.3 in which the observations were made with differing precisions. We will not do this however because it will be much more efficient to introduce matrix notation and treat the general problem of the linear estimation of  $n$  parameters using  $m$  observations subject to an aprior covariance matrix on the errors of the observations.

#### 4. The multidimensional linear estimation problem

##### 4.1. Notation and concepts of linear algebra

The anticipated applications of this paper all involve real numbers and thus we shall limit our discussion to this context. One should be aware, however, that the concepts presented have identical or very similar analogues in complex  $n$ -space. For further details on anything introduced in this section see [Golub and Van Loan] or [Lawson and Hanson].

We shall use  $R^n$  to denote  $n$ -dimensional real space. A point in  $R^n$  is an  $n$ -dimensional real vector, and will be denoted by a lower case roman or greek letter, e.g.  $x$ , with real components,  $x_1, \dots, x_n$ . An  $m \times n$  matrix is an array of  $m$  rows and  $n$  columns of real numbers, and will be denoted by an upper case roman or greek letter, e.g.  $B$ . The transpose of a matrix,  $B$ , will be denoted by  $B^t$ .

A transformation between two vector spaces will be called a linear transformation if it involves just a matrix multiplication, and an affine transformation if it consists of a matrix multiplication plus an additive constant vector.

The number of linearly independent rows of a matrix,  $B$ , is the same as the number of linearly independent columns and this number is called the rank of  $B$ . If  $B$  is  $n \times n$  and of rank  $n$  it is nonsingular and has a unique inverse matrix that we denote by  $B^{-1}$ . Also, if  $B$  is nonsingular, the matrices  $(B^t)^{-1}$  and  $(B^{-1})^t$  exist and are equal and will be denoted by  $B^{-t}$ .

The largest rank possible for an  $n \times m$  matrix is  $\min(m, n)$ . A matrix having this maximal rank is said to be of full-rank. A matrix whose rank is less than this maximal rank is called rank-deficient.

When it is necessary to distinguish between a row vector and a column vector, we shall, for example, let  $x$  denote a column vector and  $x^t$  denote a row vector. For example, if  $x$  and  $y$  are  $n$ -dimensional vectors,  $x^t y$  denotes the scalar valued inner product and  $xy^t$  denotes the  $n \times n$  matrix valued outer product.

The Euclidean norm of a vector  $x$  is denoted by

$$\|x\| = (x^t x)^{1/2}$$



The spectral norm of a matrix  $B$  is denoted by

$$\begin{aligned}\|B\| &= \text{Max}(\|Bx\| : \|x\| = 1) \\ &= [\lambda_{\max}(B^t B)]^{1/2}\end{aligned}$$

where  $\lambda_{\max}(B^t B)$  denotes the largest eigenvalue of  $B^t B$ .

Two  $n$ -vectors,  $x$  and  $y$ , are mutually orthogonal if  $x^t y = 0$ . A set of  $n$ -vectors,  $x_1, \dots, x_k$ , are mutually orthogonal if each pair is mutually orthogonal. A set of  $n$ -vectors,  $x_1, \dots, x_k$ , is orthonormal if the set is orthogonal and each vector has unit euclidean norm.

A square matrix  $Q$  is called orthogonal if its transpose is also its inverse, i.e.,  $Q^t Q = I$ , where  $I$  denotes the identity matrix. If  $Q$  is orthogonal its row vectors constitute an orthonormal set of vectors and the same is true for its column vectors.

Multiplication of a vector (respectively matrix) by an orthogonal matrix preserves its euclidean norm (respectively spectral norm). Thus if  $Q$  is an orthogonal matrix then  $\|Qx\| = \|x\|$  and  $\|QB\| = \|B\|$ .

A 2-dimensional orthogonal matrix is either a rotation matrix

$$Q = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}$$

or a reflection matrix

$$Q = \begin{bmatrix} \cos \theta & \sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}$$

An  $n$ -dimensional orthogonal matrix, with  $n \geq 2$ , can be represented as the product of at most  $n(n-1)/2$  special orthogonal matrices each of which represents either a rotation or a reflection in the plane defined by some pair of coordinate axes.

A square matrix  $A$  is symmetric if  $A^t = A$ . A symmetric matrix is positive definite if  $x^t A x > 0$  for every  $n$ -vector  $x \neq 0$  and nonnegative definite (also called positive semidefinite) if  $x^t A x \geq 0$  for every  $n$ -vector  $x \neq 0$ . Note that the class of nonnegative definite matrices includes the class of positive definite matrices. A positive definite matrix is nonsingular, and its inverse matrix is positive definite.

If  $A$  is positive definite, the scalar quantity,  $x^t A y$ , may be regarded as a generalized inner product relative to the matrix,  $A$ . The generalized inner product has analogous properties to the ordinary inner product, but with some changes of terminology. For example, whereas  $x$  and  $y$  are mutually orthogonal if  $x^t y = 0$ , they are mutually conjugate with respect to  $A$  if  $x^t A y = 0$ .

Every matrix  $A$  of the form  $A = B^t B$  or  $A = B^t W B$ , where  $B$  is any  $m \times n$  matrix and  $W$  is  $n \times n$  nonnegative definite, is nonnegative definite. If the column vectors of  $B$  are linearly independent and  $W$  is positive definite, then  $A$ , given by either of the above two expressions, is positive definite.

A partial converse of this is the fact that any  $n$ -dimensional symmetric nonnegative definite matrix,  $A$ , has a Cholesky factorization of the form,  $A = U^t U$ , where  $U$  is an  $n \times n$  upper triangular matrix. If  $A$  is positive definite, then  $U$  is uniquely determined by  $A$  to within the signs of the rows of  $U$ . That is, if  $U$  satisfies  $A = U^t U$ , then so does the matrix  $V$  obtained by multiplying any row of  $U$  by  $-1$ . Thus, if  $A$  is positive definite, we may standardize its upper triangular Cholesky factor,  $U$ , by requiring that its diagonal elements be positive, then  $U$  is uniquely determined by  $A$ .

It is sometimes more convenient to focus attention on the left member of the Cholesky factorization. Thus writing  $L$  for  $U^t$  we may write the factorization as  $A = LL^t$ .

Every symmetric matrix  $A$  has an eigensystem factorization of the form,  $A = \Lambda V V^t$ , where  $\Lambda$  is an  $n \times n$  diagonal matrix, and  $V$  is an  $n \times n$  orthogonal matrix. The diagonal elements of  $\Lambda$  are the eigenvalues of  $A$  and the column vectors of  $V$  are the eigenvectors of  $A$ . Note that the equation satisfied by these matrices can also be written as  $AV = V\Lambda$ .

The  $n$  eigenvalues of a symmetric matrix are uniquely determined by the matrix. The eigenvalues of a symmetric matrix are all positive if and only if the matrix is positive definite and are all nonnegative if and only if the matrix is nonnegative definite.

Every  $m \times n$  matrix,  $B$ , has a singular value decomposition, of the form,  $B = USV^t$ , where  $U$  is an  $m \times m$  orthogonal matrix,  $V$  is an  $n \times n$  orthogonal matrix, and  $S$  is an  $m \times n$  matrix that is all zero except for the diagonal terms, which may be positive or zero. Denoting the diagonal terms of  $S$  by  $s_i$ ,  $i = 1, \dots, \min(m, n)$ , it is often useful to assume these are ordered so that  $s_1 \geq s_2 \geq \dots$ . The numbers,  $s_i$ , are called the singular values of  $B$ . The number, say  $k$ , of nonzero singular values is equal to the rank of  $B$ . Since  $B^t B = V S^t S V^t$ , it follows that the numbers  $s_i^2$ ,  $i = 1, \dots, k$ , are the nonzero eigenvalues of  $B^t B$ , and the column vectors of  $V$  are the corresponding eigenvectors of  $B^t B$ .

The condition number of a full-rank matrix is the ratio between its largest and smallest nonzero singular values. Loosely speaking the condition number of a matrix is an upper bound on the amount by which relative errors in a vector will be magnified when the vector is operated upon by the matrix, either by direct multiplication or by solving a system. A matrix is called well-conditioned if its condition number is near one, and ill-conditioned if its condition number is large. A matrix has the minimal possible condition number of one if and only if either its rows or columns (or both) are mutually orthogonal. Thus a square matrix has a condition number of one if and only if it is an orthogonal matrix.

Every  $m \times n$  matrix,  $B$ , has a QR factorization, of the form,  $B = QR$ , where  $Q$  is an  $n \times n$  orthogonal matrix and  $R$  is an upper triangular  $m \times n$  matrix. If  $m > n$  and  $\text{Rank}(B) = n$ , the first  $n$  column vectors of  $Q$  form an orthogonal basis for the linear space spanned by the column vectors of  $B$ , and the matrix  $R$  of the QR factorization of  $B$  is also a right Cholesky factor of the positive definite matrix,  $B^t B$ , i.e.,  $B^t B = R^t R$ .

The determinant of a triangular matrix,  $L$ , denoted by  $\text{Det}(L)$ , is the product of the diagonal elements of  $L$ . The determinant of a symmetric  $n \times n$  matrix,  $A$ , is the product of the  $n$  eigenvalues of  $A$ . The determinant of the identity matrix, or of any other orthogonal matrix, is 1. The determinant of the product of a set of matrices is the product of the determinants of the matrices. As examples, if  $A$  is positive definite, with the Cholesky factorization,  $A = LL^t$ , then

$$[\text{Det}(A)]^{1/2} = \text{Det}(L)$$

and if  $B$  is a square matrix with QR factorization,  $B = QR$ , then

$$\text{Det}(B) = \text{Det}(R)$$

#### 4.2 N-dimensional random variables

An  $n$ -dimensional frequency function or probability density function is a function defined over  $R^n$  that is nonnegative and whose integral over all of  $R^n$  exists, and has the value 1. Letting  $x$  denote an  $n$ -dimensional vector, the mean value of an  $n$ -dimensional distribution having frequency function  $f$  is the  $n$ -vector  $\mu$  defined by

$$\mu = E(x) = \int x f(x) dx$$

where here the integral sign denotes integration over all of  $R_n$ .

The  $n \times n$  covariance matrix of an  $n$ -dimensional distribution with frequency function,  $f$ , is defined by

$$\begin{aligned} H = \text{Cov}(x) &= E[(x - \mu)(x - \mu)^t] \\ &= \int (x - \mu)(x - \mu)^t f(x) dx \end{aligned}$$

From the form of this expression it can be shown that the matrix  $H$  is symmetric, i.e.,  $h_{ij} = h_{ji}$  for all  $i$  and  $j$ , and also  $H$  is nonnegative definite.

It can be verified that

$$\text{Cov}(x) = E(xx^t) - \mu\mu^t$$

If a new random variable,  $u$ , is defined as a linear transformation of  $x$ , say,  $u = Ax$ , then

$$E(u) = A E(x)$$

and

$$\text{Cov}(u) = A \text{Cov}(x) A^t$$

An important special case arises when the function  $f(x)$  is the product of  $n$  functions, each depending on just one component of  $x$ , i.e.

$$f(x) = f_1(x_1) \cdots f_n(x_n)$$

In such a case we may assume without loss of generality that the functions,  $f_i$ , are scaled so that each one is a frequency function. Then the off-diagonal terms of  $H$  are all zero, and each diagonal term,  $h_{ii}$ , is just the one-dimensional variance of the component,  $x_i$ , determined by the frequency function  $f_i(x_i)$ .

The separate components of  $x$  are said to be independently distributed if and only if all of the off-diagonal elements of the covariance matrix are zero.

#### 4.3. Linear estimation of $n$ parameters using $m$ observations

Assume that an  $m$ -dimensional phenomenon (this may be a number of instances of some lower dimensional phenomena) to be observed has a distribution,  $D$ , with a frequency function,  $f$ , having an  $m$ -dimensional mean vector,  $\eta$ , and an  $m \times m$  positive definite covariance matrix,  $H$ .

Assume further that  $\eta$  is representable as a linear combination of  $n$   $m$ -vectors,  $b_i$ ,  $i = 1, \dots, n$ . Thus we are assuming there are coefficients,  $\xi_i$ , such that

$$\eta = \sum_i \xi_i b_i$$

Letting  $B$  denote the  $m \times n$  matrix with column vectors  $b_i$ , and  $\xi$  denote the  $n$ -vector with components  $\xi_i$ , this equation can be written as

$$\eta = B\xi$$

To avoid complications that would obscure the central ideas, we assume that  $m > n$ , and the vectors  $b_i$  are linearly independent. It follows that  $B$  is of rank  $n$ .

Various linear estimation problems can be based on this model, depending on which elements of the model are assumed to be known and which are to be estimated. Consider first the case in which  $H$  and  $B$  are known, and  $\xi$  and  $\eta$  are unknown. Suppose we have an observation,  $y$ , regarded as a random sample from the distribution,  $D$ . We wish to estimate  $\xi$  and  $\eta$  and the covariance matrices of each of these estimates.

The Method of presentation used in Sections 4.3.1 through 4.3. follows pp. 36-48 of [Plackett, 1960].

##### 4.3.1. The estimator $\hat{\xi}$ and its covariance matrix

First note that  $y$  itself is a linear unbiased estimator for  $\eta$  since  $E(y) = \eta$ , however we shall see that a better estimator is available.

Our linear estimator for  $\xi$  will be

$$(1) \quad \hat{\xi} = (B^t H^{-1} B)^{-1} B^t H^{-1} y = P y$$

Here we have introduced the  $n \times m$  matrix

$$(2) \quad P = (B^t H^{-1} B)^{-1} B^t H^{-1}$$

The estimator  $\hat{\xi}$  is clearly the unique solution of the linear system

$$(3) \quad B^t H^{-1} B \hat{\xi} = B^t H^{-1} y$$

which will be discussed further in Sec. 4.3.3.

To verify that  $\hat{\xi}$  is an unbiased linear estimator for  $\xi$  we compute

$$\begin{aligned} E[\hat{\xi}] &= E[Py] = P E[y] = P \eta = P B \xi \\ &= [(B^t H^{-1} B)^{-1} B^t H^{-1}] B \xi \\ &= (B^t H^{-1} B)^{-1} (B^t H^{-1} B) \xi \\ &= \xi \end{aligned}$$

Note that the property of  $P$  that was crucial here was

$$PB = I$$

Since  $B$  is a rectangular matrix it does not have an inverse, but any matrix  $G$  satisfying  $GB = I$  is called a left inverse of  $B$ . There will in general be many such matrices and all provide unbiased linear estimators for  $\xi$ .

The covariance matrix for the estimator,  $\hat{\xi}$ , may be computed as

$$\begin{aligned} \text{Cov}(\hat{\xi}) &= \text{Cov}(Py) = P \text{Cov}(y) P^t = P H P^t \\ &= [(B^t H^{-1} B)^{-1} B^t H^{-1}] H [H^{-1} B (B^t H^{-1} B)^{-1}] \\ &= (B^t H^{-1} B)^{-1} (B^t H^{-1} B) (B^t H^{-1} B)^{-1} \\ &= (B^t H^{-1} B)^{-1} \end{aligned}$$

Among unbiased linear estimators, the estimator,  $Py$ , has the minimum variance, in the sense that if  $Gy$  is any other unbiased linear estimator, i.e.,  $G$  satisfies  $GB = I$ , then the difference,  $\text{Cov}(Gy) - \text{Cov}(Py)$ , will be a nonnegative definite matrix. To verify this we write a matrix expression that is nonnegative definite due to its form and then show it is equal to the required difference.

$$\begin{aligned} (G-P)H(G-P)^t &= GHG^t - GHP^t - PHG^t + PHP^t \\ &= \text{Cov}(Gy) - GH[H^{-1}B(B^t H^{-1}B)^{-1}] \\ &\quad - [(B^t H^{-1}B)^{-1} B^t H^{-1}] HG^t + \text{Cov}(Py) \\ &= \text{Cov}(Gy) - \text{Cov}(Py) - \text{Cov}(Py) + \text{Cov}(Py) \end{aligned}$$

$$= \text{Cov}(Gy) - \text{Cov}(Py)$$

#### 4.3.2. The estimator $\hat{\eta}$ and its covariance matrix

Since  $\eta = B\xi$ , we define our estimator for  $\eta$  to be

$$\hat{\eta} = B\hat{\xi} = BPY$$

This estimator is unbiased since

$$E[\hat{\eta}] = B E[\hat{\xi}] = B \xi = \eta$$

The covariance matrix for  $\hat{\eta}$  is

$$\text{Cov}(\hat{\eta}) = \text{Cov}(B\hat{\xi}) = B \text{Cov}(\hat{\xi}) B^t$$

Since we obtained two different expressions for  $\text{Cov}(\hat{\xi})$  we may write either

$$\text{Cov}(\hat{\eta}) = B PHP^t B^t = BP H (BP)^t$$

or

$$\text{Cov}(\hat{\eta}) = B (B^t H^{-1} B)^{-1} B^t$$

Since  $P$  is a left inverse for  $B$ , it follows that  $BP$  is a left identity for  $B$ , i.e.,  $(BP)B = B$ . It can be verified that this is the crucial property that permitted verification that  $BPY$  is an unbiased estimator of  $\eta$ . Thus if  $J$  is any left identity for  $B$ , i.e.,  $J$  is an  $m \times n$  matrix satisfying  $JB = B$ , then  $Jy$  is an unbiased estimator for  $\eta$ .

Among all unbiased linear estimators for  $\eta$ ,  $BPY$  has the minimum variance, in the sense that if  $Jy$  is any other unbiased linear estimator, i.e.,  $J$  satisfies  $JB = B$ , then the difference,  $\text{Cov}(Jy) - \text{Cov}(BPY)$ , is a nonnegative definite matrix. This is verified as follows:

$$\begin{aligned} (J - BP)H(J - BP)^t &= J H J^t - J H P^t B^t - BP H J^t + BP H (BP)^t \\ &= \text{Cov}(J) - J H [H^{-1} B (B^t H^{-1} B)^{-1}] B^t \\ &\quad - B [(B^t H^{-1} B)^{-1} B^t H^{-1}] H J^t + \text{Cov}(BP) \\ &= \text{Cov}(J) - B (B^t H^{-1} B)^{-1} B^t \\ &\quad - B (B^t H^{-1} B)^{-1} B^t + \text{Cov}(BP) \\ &= \text{Cov}(J) - \text{Cov}(BP) \end{aligned}$$

#### 4.3.3. Interpretation as least squares estimation

The estimator,  $\hat{\xi}$ , defined in Eq. 4.3.1.(1), can be derived as the solution to a certain weighted linear least squares problem. This is the problem of finding an  $n$ -vector,  $x$ , to minimize the quadratic function

$$(1) \quad S^2 = (Bx - y)^t H^{-1} (Bx - y)$$

where  $B$  is a given  $m \times n$  matrix,  $y$  is an  $m$ -vector of observations, and  $H$  is an  $m \times m$  positive definite symmetric matrix which is the known covariance matrix for the distribution of errors in  $y$ .

The  $n$ -vector of partial derivatives of  $S^2$  with respect to  $x$  is

$$(2) \quad \partial S^2 / \partial x = 2B^t H^{-1} (Bx - y)$$

Setting this equal to zero gives the system of equations

$$(3) \quad B^t H^{-1} Bx = B^t H^{-1} y$$

to be solved for  $x$ . We shall let  $\hat{S}^2$  denote the minimum value of  $S^2$ , i.e., the value of  $S^2$  when  $x = \hat{x}$ .

Eq.(3) (see also Eq. 4.3.1.(3)) is called the normal equations for the least squares problem of minimizing  $S^2$  of Eq.(1). Note that when  $H = I$ , the geometric interpretation of setting Eq.(2) equal to zero is that the residual vector,  $Bx - y$  is required to be orthogonal (i.e., perpendicular or normal) to all columns of the matrix,  $B$ . I presume this is the reason the word "normal" has been associated with Eq.(3). When  $H \neq I$ ,  $Bx - y$  is required to be conjugate to all columns of  $B$ , relative to the positive definite matrix,  $H^{-1}$ .

The use of normal equations in the form  $B^t Bx = B^t y$  dates back to Gauss, 1821. The form treated here, involving  $H$ , was introduced by A. C. Aitken, 1934. Reference: [Plackett].

#### 4.3.3.1. Transformations of the least squares problem

There are two types of transformations that are very useful, both for the analysis and for the computational solution of a least squares problem. The first transforms the case of general  $H$  to the case of  $H = I$ , while the second decomposes  $m$ -space into the  $n$ -dimensional subspace spanned by the columns of  $C$ , and the complementary  $(m-n)$ -dimensional subspace orthogonal to the columns of  $C$ .

For the first transformation we factor  $H$  as

$$(1) \quad H = LL^t$$

and introduce  $C$  and  $z$  defined by

$$(2) \quad C = L^{-1}B$$

and

$$(3) \quad z = L^{-1}y$$

Then Eq. 4.3.3.(1) can be rewritten as

$$(4) \quad S^2 = (Cx - z)^t (Cx - z) = \|Cx - z\|^2$$

The statistical interpretation of this transformation is that  $y$ , which was a sample from a distribution having mean,  $\eta$ , and covariance matrix,  $H$ , is transformed to  $z$ , which is a sample from a distribution having mean,  $L^{-1}\eta$ , and covariance matrix,  $I$ . Thus the components of  $z$  each have unit variance and are mutually uncorrelated.

There are various ways to achieve the factorization,  $H = LL^t$ . The simplest is the Cholesky decomposition of  $H$ . Another approach would be to use the eigensystem decomposition,  $H = V\Lambda V^t$ , and then set  $L = V\Lambda^{1/2}$ .

The goal of the second transformation is to replace  $C$  by a matrix having all of its nonzero elements in its first  $n$  rows. It is convenient, but not essential, to choose this transformation so that it also transforms  $z$  to a vector having all of its nonzero components in the first  $n+1$  positions.

One way to construct such a transformation is by use of the QR decomposition of  $C$ ,

$$(5) \quad C = Q \begin{bmatrix} R \\ 0 \end{bmatrix}$$

where  $Q$  is an  $m \times n$  orthogonal matrix,  $R$  is an  $n \times n$  nonsingular upper triangular matrix, and  $0$  denotes a zero matrix of conformable dimensions, here an  $(m-n) \times n$  zero matrix. (An alternative to the use of the QR decomposition for this transformation would be the use of the singular value decomposition. The SVD is computationally more expensive, but gives additional information that is desirable in some situations. We shall not discuss the SVD further in this paper.)

Using Eq.(5) in Eq.(4) gives

$$(6) \quad S^2 = \left\| Q \begin{bmatrix} R \\ 0 \end{bmatrix} x - z \right\|^2 = \left\| \begin{bmatrix} R \\ 0 \end{bmatrix} x - Q^t z \right\|^2$$

Let  $u$  denote the first  $n$  components of  $Q^t z$  and let  $v$  denote the last  $m-n$  components of  $Q^t z$ , i.e.,

$$(7) \quad \begin{bmatrix} u \\ v \end{bmatrix} = Q^t z$$

Then

$$(8) \quad S^2 = \left\| \begin{bmatrix} R \\ 0 \end{bmatrix} x - \begin{bmatrix} u \\ v \end{bmatrix} \right\|^2 = \|Rx - u\|^2 + \|v\|^2$$

This last expression shows  $S^2$  as the sum of two terms, the first of which can be reduced to zero by the unique  $x$  that satisfies  $Rx = u$ , while the second term is independent of  $x$ . Thus the minimum value of  $S^2$  is

$$(9) \quad S^2 = \|v\|^2$$

and since the minimizing value of  $x$  is unique, and is already known to be  $\hat{\xi}$ , it follows that  $\hat{\xi}$  satisfies



$$(10) \quad R\hat{\xi} = u$$

From  $\text{Cov}(z) = I$ , Eq.(7), and the orthogonality of  $Q$ , we obtain

$$(11) \quad \text{Cov}\left(\begin{bmatrix} u \\ v \end{bmatrix}\right) = I$$

and thus, also

$$(12) \quad \text{Cov}(u) = I$$

and

$$(13) \quad \text{Cov}(v) = I$$

From Eq.(9) we obtain

$$(14) \quad E(\hat{S}^2) = E(\|v\|^2) = \sum E(v_i^2) = \sum \text{Var}(v_i) = m-n$$

#### 4.3.4. Introducing a scaling factor for $H$

The assumption that  $H$  is completely known a priori is often unrealistic. A useful weakening of this assumption is the assumption that the covariance matrix of the distribution,  $D$ , from which the observation was sampled, is  $\phi^2 H$ , where  $H$  is a known positive definite matrix and  $\phi$  is an unknown scalar. This amounts to assuming that the signs and relative sizes of the elements of the covariance matrix of  $D$  are known but an overall scale factor is unknown. We shall see that this factor,  $\phi^2$ , can be estimated using the sum of squares of residuals,  $S^2$ , of Eq. 4.3.3.(1), generalizing the example discussed in Sec. 3.

As to practical methods for the a priori definition of  $H$ , a common situation would be to assume the errors in the different components of  $y$  are uncorrelated so that  $H$  need only be a diagonal matrix. Then each diagonal element of  $H$  would be assigned the a priori variance of the error in the corresponding component of  $y$ . In this case one would expect the value of  $\phi$  to turn out to be 1, and the extent to which the a posteriori estimate of  $\phi$  differs from 1 can be a consideration in assessing the quality of the model relative to the available data.

Another possibility is that one may simply assume that the errors in the different components of  $y$  are uncorrelated and all have the same, but unknown, variance. Then one could set  $H = I$  and the a posteriori estimated value of  $\phi^2$  would be an estimate of the variance of errors for each component of  $y$ .

If we start with  $\phi^2 H$  as the covariance matrix for  $D$  and repeat the derivations of Sections 4.3.1, 4.3.2, 4.3.3, and 4.3.3.1, we find  $\phi^2$  cancels out in the expression for the matrix,  $P$ , so the estimator,  $\hat{\xi}$ , is unchanged. The covariance matrix for  $\hat{\xi}$  inherits the factor  $\phi^2$ , so we obtain

$$(1) \quad \text{Cov}(\hat{\xi}) = \phi^2 (B^t H^{-1} B)^{-1}$$

Similarly the estimator  $\hat{\eta}$  is unchanged, but its covariance becomes

$$(2) \quad \text{Cov}(\hat{\eta}) = \phi^2 B P H (BP)^t$$

or

$$(3) \quad \text{Cov}(\hat{\eta}) = \phi^2 B (B^t H^{-1} B)^{-1} B^t$$

Eqs. 4.3.3.1. (11-14) become

$$(4) \quad \text{Cov} \begin{pmatrix} u \\ v \end{pmatrix} = I$$

$$(5) \quad \text{Cov}(u) = I$$

$$(6) \quad \text{Cov}(v) = I$$

$$(7) \quad E(\tilde{S}^2) = (m \cdot n) \phi^2$$

#### 4.3.4.1. An unbiased estimator for $\phi$

Eq. 4.3.4.(7) can be rewritten as

$$(1) \quad E[\tilde{S}^2/(m \cdot n)] = \phi^2$$

showing that  $\tilde{S}^2/(m \cdot n)$  is an unbiased estimator for  $\phi^2$ . We denote this estimator by

$$(2) \quad \hat{\phi}^2 = \tilde{S}^2/(m \cdot n)$$

This provides a practical estimate for  $\phi$  that is needed in situations such as were described at the beginning of Sec. 4.3.4.

The next question one typically asks is how good is this estimate. In terms of our mathematical model this translates to the problem of finding the variance of the estimator,  $\hat{\phi}$ .

A direct derivation of this variance for the case of a general underlying frequency function,  $f$ , would involve third and fourth moments of  $f$ , i.e. integrals of  $y^3 f(y)$  and  $y^4 f(y)$ . This would not be of practical use since independent estimates of these moments would not generally be known. Thus, as was mentioned in Sec. 3, the usual approach is to assume that  $f$  is some well known frequency function that provides a plausible model for the observation errors in the real world system being investigated, and for which the resulting distribution for  $\tilde{S}^2$  is known.

Standard statistical distributions of interest for our purposes will be presented in Sec. 5.

## 5. Standard distributions

### 5.1. The normal or Gaussian distribution

The one-dimensional normal distribution, also called the Gaussian distribution, having mean,  $\eta$ , and variance,  $\sigma^2$ , is conventionally denoted by  $N(\eta, \sigma^2)$ . Its frequency function is

$$f(y) = (2\pi)^{-1/2} \sigma^{-1} \exp\{-(y-\eta)/\sigma\}^2/2\}$$

If  $y$  has the distribution,  $N(\eta, \sigma^2)$ , then the random variable defined by  $z = (y-\eta)/\sigma$  has the distribution,  $N(0,1)$ . Values of the frequency function and distribution function for  $N(0,1)$  are readily available in tables and from computer subroutines. Some probabilities from the distribution,  $N(0,1)$ , are given in the following table:

Table 1. Probabilities for  $z \in N(0,1)$

$\rho$	$P( z  \leq \rho)$
0.5	0.4
0.52	0.40
0.68	0.50
0.84	0.60
1.0	0.683
1.96	0.95
2.0	0.954
3.0	0.997
$+\infty$	1.0

Let  $N(m; \eta, \Sigma)$  denote the  $m$ -dimensional normal distribution having mean vector,  $\eta$ , and covariance matrix,  $\Sigma$ . Here  $\eta$  is an  $m$ -vector and  $\Sigma$  is a positive definite  $m \times m$  matrix. The frequency function for this distribution is

$$f(y) = (2\pi)^{-m/2} [\text{Det}(\Sigma)]^{-1/2} \exp\{-(y-\eta)^t \Sigma^{-1} (y-\eta)/2\}$$

A very significant property of the normal distribution is the fact that an affine transformation of a normal random variable will again be a normal variable. Specifically let  $P$  be an  $n \times m$  matrix with  $n \leq m$  and  $\text{Rank}(P) = n$ . Let  $y \in N(m; \eta, \Sigma)$ . Define a new  $n$ -dimensional random variable,  $z = \zeta + Py$ . Then  $z \in N(n; \zeta + P\eta, P\Sigma P^t)$ .

As an important special case of such an affine transformation, let  $E_i$  denote the  $1 \times m$  matrix (i.e. row vector) whose elements are all zero except for a 1 in column  $i$ . The product  $E_i y$  is just the component  $y_i$  of  $y$ , and  $E_i \Sigma E_i^t$  is just the component  $\sigma_{ii}$  of  $\Sigma$ . It follows that  $y_i \in N(\eta_i, \sigma_{ii})$ . [The notation here is a bit unfortunate. Note that  $\sigma_{ii}$  is the variance of  $y_i$  and thus the standard deviation of  $y_i$  is  $\sigma_{ii}^{1/2}$ ].

An affine transformation that is often useful is the transformation from the general case of  $N(m; \eta, \Sigma)$  to the special case of  $N(m; 0, I)$ . Denote the Cholesky factorization of  $\Sigma$  by  $\Sigma = LL^t$ . Then if  $y$  has the distribution,  $N(m; \eta, \Sigma)$ , the random variable defined by

$$z = L^{-1}(y-\eta)$$

has the distribution,  $N(m;0,I)$ .

As an example of the multivariate normal distribution, consider  $N(2;0,I)$ , which is also called the circular bivariate normal distribution. The integral of the density function for this distribution over a disk of radius,  $\rho$ , has a particularly simple expression, namely,  $1-\exp(-\rho^2/2)$ . Some probabilities from this distribution are given in the following table:

Table 2. Probabilities for  $z \in N(2;0,I)$

$\rho$	$\rho^2$	$P(\ z\ ^2 \leq \rho^2) = 1-\exp(-\rho^2/2)$
0.32	0.10	0.050
1.00	1.00	0.393
1.18	1.39	0.500
2.00	4.00	0.865
2.45	5.99	0.950
3.00	9.00	0.989
$+\infty$	$+\infty$	1.000

## 5.2. Confidence regions

A region in which a random sample is expected to appear  $p\%$  of the time is called a  $p\%$  confidence region. For example, if  $z \in N(0,1)$  the interval  $[-0.84, 0.84]$  is a 60% confidence interval for  $z$ . There are infinitely many other 60% confidence regions, however, for example,  $[-0.05, +\infty]$ , or  $[-\infty, 0.5]$ , or  $[-\infty, -0.52] \cup [0.52, +\infty]$ .

In higher dimensional spaces there is even more variety in the shapes and connectedness properties of regions that could be chosen to be a  $p\%$  confidence region. For example one could choose a rectangle, or other polygon, or a circle, or nonconvex or disjoint regions. A choice having practical appeal for our purposes is the one that can be characterized as being the region in which the density function exceeds a certain fixed value. Then the probability density is larger at every point interior to the region than it is at any point exterior to the region. In the case of the multivariate normal distribution such regions will be ellipsoids. Recall that spheres, ellipses, circles, and one dimensional intervals are all special cases of ellipsoids.

As we shall see in the immediately following paragraphs, this choice of the definition of a confidence region provides a significant technical convenience, in that the case of a general ellipsoid reduces easily to the case of a sphere, and thence to dependence on a scalar, rather than an  $n$ -dimensional, random variable.

For  $y \in N(m;\eta,\Sigma)$ , we shall define the  $p\%$  confidence ellipsoid to be the unique ellipsoid,  $C$ , defined by

$$(1) \quad C = \{y : (y-\eta)^t \Sigma^{-1}(y-\eta) \leq \rho^2\}$$

where  $\rho$  is the unique nonnegative value that permits the integral of the density function over  $C$  to have the value,  $p/100$ .

Note that the frequency function for the  $n$ -dimensional normal distribution is the product of  $[\text{Det}(\Sigma)]^{-1/2}$  and a function of  $(y-\eta)^t \Sigma^{-1}(y-\eta)$  with  $\Sigma$  being positive definite. Any function of this form has the property that for a fixed dimension,  $n$ , its integral over an ellipsoid  $C$  defined as in Eq.(1) depends only on  $\rho$  and not on  $\Sigma$  or  $\eta$ . Thus the functional relation between  $\rho$  and  $p$  given in Table 2 applies not only to the circular bivariate normal distribution but also to the general (elliptical) bivariate normal distribution with  $\|z\|^2$  in the table heading replaced by  $(y-\eta)^t \Sigma^{-1}(y-\eta)$ .

We see that determining critical sizes for confidence ellipses defined as in Eq.(1) depends on knowledge of the distribution of  $\|z\|^2$ , for  $z \sim N(m; 0, 1)$ . This distribution is called the  $\chi^2$  distribution with  $m$  degrees of freedom, and will be discussed in Sec. 5.3.

#### 5.2.1. Geometric characterization of a confidence ellipsoid

Let  $C$  be the ellipsoid defined above in terms of  $\eta$ ,  $\Sigma$ , and  $\rho$ . Write the eigenvalue factorization of  $\Sigma$  as  $\Sigma = Q\Lambda Q^t$  where  $Q$  is orthogonal and  $\Lambda$  is diagonal with positive diagonal elements (the eigenvalues of  $\Sigma$ ),  $\lambda_i$ ,  $i = 1, \dots, m$ . We shall assume the eigenvalues are ordered so that  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$ . The column vectors of  $Q$ , which we shall denote by  $q_i$ ,  $i = 1, \dots, m$ , are the corresponding eigenvectors of  $\Sigma$ .

Note that  $\Sigma^{-1} = Q\Lambda^{-1}Q^t$ , so the eigenvectors of  $\Sigma^{-1}$  are the same as those of  $\Sigma$  while the eigenvalues are the reciprocals of those of  $\Sigma$ .

The ellipsoid,  $C$ , is centered at  $\eta$  and its principal axes are parallel to the eigenvectors of  $\Sigma$  with  $q_1$  giving the direction of the major axis. The semi-diameter,  $\alpha_i$ , in the direction of the  $i^{\text{th}}$  principal axis can be determined by solving the equation

$$(\alpha_i q_i)^t \Sigma^{-1} (\alpha_i q_i) = \rho^2$$

from which

$$\alpha_i^2 \lambda_i^{-1} = \rho^2$$

$$\alpha_i = \rho \lambda_i^{1/2}$$

Consider the case in which  $\Sigma = \sigma^2 I$ . Then the eigenvalues of  $\Sigma$  are all equal to  $\sigma^2$ , so the confidence ellipsoid is a sphere of radius  $\alpha = \rho\sigma$ .

#### 5.3. The $\chi^2$ distribution

We have encountered two different situations, Eq. 4.3.3.1.(14) and Sec. 5.2, in which our analysis led to consideration of the distribution of the sum of squares of random variables. If these variables are independent samples from  $N(0,1)$ , their sum of squares has a distribution called the  $\chi^2$  distribution.

Eq. 4.3.3.1.(9) showed that  $\bar{S}^2$  was the sum of squares of the components of the  $(m-n)$ -vector,  $v$ , and it was also established that  $E(v) = 0$  and  $\text{Cov}(v) = \phi^2 I$ . If we now add the assumption that the distribution of the original random variable,  $y$ , is the normal distribution, i.e.,  $y \in N(m; \eta, \phi^2 \Sigma)$ , then, since  $v$  was obtained as an affine transformation of  $y$ , the distribution of  $v$  is  $N(n; 0, \phi^2 I)$ .

A random variable defined as the sum of squares of  $k$  independent samples from  $N(0,1)$ , or equivalently as the sum of squares of components of a sample vector from  $N(k; 0, I)$ , is said to have the  $\chi^2$  (chi-squared) distribution with  $k$  degrees of freedom. We will denote this distribution by  $\chi^2[k]$ .

Letting  $t$  denote the independent variable, the density function for the  $\chi^2[k]$  distribution is zero for  $t < 0$  and

$$f(t) = \frac{t^{(k/2)-1} e^{-t/2}}{2^{k/2} \Gamma(k/2)} \quad \text{for } t \geq 0$$

The distribution  $\chi^2[k]$  has a mean value of  $k$ . The variance is  $2k$  and thus the standard deviation is  $(2k)^{1/2}$ . Values of the distribution function for  $\chi^2[k]$  are available from tables or computer subroutines.

Unlike the normal distribution, the  $\chi^2$  distribution is unsymmetric. The mode and median of the  $\chi^2[k]$  distribution are each less than the mean. If  $s$  is a  $\chi^2[k]$  variable, the transformed variable,  $t = (2s)^{1/2}$ , has a distribution that is closely approximated by  $N((2k-1)^{1/2}, 1)$  for  $k \geq 30$ .

Table 3. Values of  $P(s \leq x)$  for  $s \in \chi^2[k]$  for selected values of  $k$  and  $x$ .

$k$	$x =$	0.05	0.5	0.95
1		0.004	0.46	3.8
2		0.10	1.39	5.99
5		1.1	4.4	11.1
10		3.9	9.3	18.3
20		10.9	19.3	31.4
30		18.5	29.3	43.8

Note that the row for  $k = 2$  in this table agrees with values in Table 2 since  $\|z\|^2$  of Table 2 is a  $\chi^2[2]$  variable.

We may now treat the question raised in Sec. 4.3.4 regarding the dispersion of  $\bar{S}^2/(m-n)$  as an estimator for  $\phi^2$ . If we assume the data vector  $y$  is a sample from  $N(m; \eta, \phi^2 H)$ , with  $H$  known, then  $\bar{S}^2/\phi^2$  is a sample from  $\chi^2[m-n]$ .

For example, if  $m-n = 20$ , we may conclude that there is a 90% probability that

$$10.9 \leq \bar{S}^2/\phi^2 \leq 31.4$$

and a 5% probability that

$$\hat{S}^2/\phi^2 \geq 31.4$$

There are various ways one may use these results. If one has a prior notion of a reasonable value for  $\phi$ , then if  $\hat{S}^2$  exceeds 31.4 times the square of that value, it may be taken as evidence that the model is not consistent with the data, or the prior value of  $\phi$  was incorrect.

If one has no prior notion of the value of  $\phi$ , then one might conclude that there is a 90% probability that  $\phi^2$  satisfies

$$\hat{S}^2/31.4 \leq \phi^2 \leq \hat{S}^2/10.9$$

#### 5.4. Student's t distribution

In Sec. 4.3.4 the formula  $\text{Cov}(\hat{\xi}) = \phi^2(B^t H^{-1} B)^{-1}$  was obtained. Also it has been noted that if  $y$  has a normal distribution then  $\hat{\xi}$  does also, in particular

$$(1) \quad \hat{\xi} \in N(n; \xi, \phi^2(B^t H^{-1} B)^{-1})$$

Thus if  $B$ ,  $H$ , and  $\phi$  are known, one can compute a  $p\%$  confidence ellipsoid for  $\xi$ , or  $p\%$  confidence intervals for individual components of  $\xi$ , as discussed in Sec. 5.2.

If  $\phi$  is not known it can be replaced by its estimator,

$$(2) \quad \hat{\phi} = [\hat{S}^2/(m-n)]^{1/2}$$

(see Eq. 4.3.4.2.(2)), if  $m-n$  is sufficiently large so the variance of  $\hat{\phi}$  is sufficiently small. Alternatively, particularly when  $m-n$  is not large, a different approach may be used, involving the Student's t distribution. This distribution was presented in 1908 by an anonymous author identified as "Student". Reference: [Plackett].

Eq.(1) can be rewritten as

$$(3) \quad \hat{\xi}/\hat{\phi} \in N(n; \xi/\phi, (B^t H^{-1} B)^{-1})$$

which serves to focus attention on  $\hat{\xi}/\hat{\phi}$ . We now ask: What is the distribution of  $\hat{\xi}/\hat{\phi}$ ?

Note that  $\hat{\xi}$  has a normal distribution and  $\hat{\phi}^2$  is within a known factor of being a  $\chi^2$  variable. Furthermore these two variables are statistically independent, since  $\hat{\xi}$  is a function of  $u$  and not  $v$ , [Eq.4.3.3.1.(10)],  $\hat{S}^2$  is a function of  $v$  and not  $u$ , [Eq.4.3.3.1.(14)], and  $u$  and  $v$  are statistically independent, [Eq.4.3.3.1.(11) or 4.3.4.(4)].

Thus we need to know the distribution of the ratio of a normal variable to the square root of a statistically independent  $\chi^2$  variable. That is what the Student's t distribution is.

Let  $w \in N(n; 0, \Sigma)$  and  $\beta^2 \epsilon \chi^2[k]$ , with these distributions being mutually independent. Then the  $n$ -dimensional vector random variable

$$(4) \quad t = k^{1/2} w / \beta$$

has the  $n$ -dimensional Student's  $t$  distribution with  $k$  degrees of freedom and kernel matrix  $\Sigma$ . The frequency function for  $t$  is

$$(5) \quad f(t) = \frac{\Gamma[(k+n)/2]}{\Gamma(k/2)(k\pi)^{n/2} [\text{Det}(\Sigma)]^{1/2} [1 + t^t \Sigma^{-1} t]^{(k+n)/2}}$$

Furthermore,

$$(6) \quad E(t) = 0$$

and

$$(7) \quad \text{Cov}(t) = [k/(k-2)] \Sigma$$

Remark: The frequency function for the 1-dimensional Student's  $t$  distribution with  $\Sigma = 1$  can be found in numerous sources, however I wish to thank Dr. J. Myhre for referring me to [Anderson] where Eqs. (5-7) above are given in Problem 27, Page 283.

We return now to the study of  $\hat{\xi}/\hat{\phi}$ . To obtain  $w$  and  $\beta$  satisfying the conditions to define an  $n$ -dimensional Student's  $t$  variable, let  $w = (\hat{\xi} - \xi)/\phi$  and  $\beta = \hat{S}/\phi$ . Then the variable

$$(8) \quad t = (m-n)^{1/2} (\hat{\xi} - \xi) / \hat{S} = (\hat{\xi} - \xi) / \hat{\phi}$$

has the Student's  $t$  distribution with  $(m-n)$  degrees of freedom and kernel matrix  $\Sigma = (B^t H^{-1} B)^{-1}$ .

As with the density function for the multivariable normal distribution, the density function for the multivariable Student's  $t$  distribution is the product of  $[\text{Det}(\Sigma)]^{-1/2}$  and a function of  $t^t \Sigma^{-1} t$ , with  $\Sigma$  being positive definite, and thus has the property that for a fixed dimension,  $n$ , and number of degrees of freedom,  $k$ , its integral over an ellipsoid  $C$  defined as in Eq. 5.2.(1) depends only on  $\rho$  and not on  $\Sigma$  or  $\eta$ .

Thus to find critical values for confidence ellipses of the form of Eq. 5.2.(1) for an  $n$ -dimensional Student's  $t$  distribution with  $k$  degrees of freedom, it suffices to know the distribution of the scalar random variable,  $\|t\|^2$ . The variable  $\|t\|^2/n$  has a distribution known as the  $F$  distribution with  $n$  and  $k$  degrees of freedom.

We shall define the  $F$  distribution in Sec. 5.5, then verify the assertion in the last sentence above, and finally relate the  $F$  distribution to the our particular case of  $t = (\hat{\xi} - \xi)/\hat{\phi}$ .



## 5.5. The F distribution

If  $\theta_1 \in \chi^2[k_1]$  and  $\theta_2 \in \chi^2[k_2]$  then the ratio

$$\psi = \frac{\theta_1/k_1}{\theta_2/k_2} = \frac{k_2\theta_1}{k_1\theta_2}$$

is said to have the F distribution with  $k_1$  and  $k_2$  degrees of freedom. This distribution will be denoted by  $F[k_1, k_2]$ .

The F distribution was introduced by Fisher, 1922, and incorporated in a general framework for testing linear hypotheses by Kolodziejczyk, 1935. Reference: [Plackett]. Values of the distribution function for the F distribution are available in tables and from computer subroutines.

Suppose  $w \in N(n; 0, I)$ ,  $\theta_2 \in \chi^2[k_2]$ , and  $t = k_2^{1/2}w/\theta_2$ . Then  $t$  is an  $n$ -dimensional Student's  $t$  variable with  $k_2$  degrees of freedom. Also  $\|w\|^2$  is a  $\chi^2[n]$  variable. Let

$$\psi = \frac{\|t\|^2}{n} = \frac{k_2 \|w\|^2}{n \theta_2^2}$$

Then  $\psi \in F[n, k_2]$ .

Continuing now our consideration of  $\hat{\xi}/\hat{\phi}$ , it was noted (See Eq. 5.4.(8)) that  $(\hat{\xi}-\xi)/\hat{\phi}$  is an  $n$ -dimensional Student's  $t$  variable with  $(m-n)$  degrees of freedom. It follows therefore that  $\|\hat{\xi}-\xi\|^2/(n\hat{\phi}^2) \in F[n, m-n]$ . Thus critical points for confidence ellipses for  $\xi$  when  $\phi$  has been estimated by  $\hat{\phi}$  can be obtained from tables for the F distribution.

## 6. Approaches to data analysis

### 6.1. Analysis of one coherent set of data

Assume we have a set of  $m$  observed values,  $y_i$ ,  $i = 1, \dots, m$ , which we will treat as an  $m$ -dimensional vector. It is assumed that  $y$  can be regarded as being a random sample from the normal distribution,  $N(m; \eta, \phi^2 H)$ , where  $H$  is a known  $m \times m$  positive definite matrix and  $\eta$  is unknown. We think that the value of  $\phi$  is 1 but we will estimate  $\phi$  from the data as a check on the validity of the model.

It is assumed that  $\eta$  has a representation of the form

$$(1) \quad \eta = B\xi$$

where  $B$  is a known  $m \times n$  matrix of rank  $n$ , and  $\xi$  is an unknown  $n$ -vector. Our objectives are to estimate  $\xi$ , obtain a covariance matrix for the estimator of  $\xi$  as a measure of the dispersion of the estimator, and estimate  $\phi$  as a measure of the quality of the model.

We shall use the two transformations introduced in Sec. 4.3.3.1. Details of the computational steps specified will be found in [Lawson and Hanson].

Compute the Cholesky factorization of  $H$  as  $H = LL^t$ . Then we know the minimum variance unbiased linear estimator,  $\hat{\xi}$ , for  $\xi$  is the value of  $x$  that solves the least squares problem of

$$(2) \quad \text{minimizing } \|L^{-1}Bx - L^{-1}y\|^2$$

Compute  $[C:z] = L^{-1}[B:y]$  by solving the system  $L[C:z] = [B:y]$ . Then our least squares problem is to

$$(3) \quad \text{minimize } \|Cx - z\|^2$$

Compute the "QR" factorization of  $[C:z]$ . This gives an  $m \times m$  orthogonal matrix,  $Q$ , and an  $(n+1) \times (n+1)$  upper triangular matrix,  $U$ , satisfying

$$(4) \quad [C:z] = Q \begin{bmatrix} U \\ 0 \end{bmatrix}$$

Partition the matrix  $U$  as

$$(5) \quad U = \begin{bmatrix} R & g \\ 0 & \alpha \end{bmatrix}$$

where  $R$  is an  $n \times n$  nonsingular upper triangular matrix,  $g$  is an  $n$ -dimensional column vector,  $0$  denotes an  $n$ -dimensional row vector of zeros, and  $\alpha$  is a scalar. The transformed least squares problem is now that of

$$(6) \quad \text{minimizing } \left\| \begin{bmatrix} R \\ 0 \end{bmatrix} x - \begin{bmatrix} g \\ \alpha \end{bmatrix} \right\|^2$$

or equivalently

$$(7) \quad \text{minimizing } \|Rx - g\|^2 + \alpha^2$$

Since  $R$  is nonsingular, the first term in this expression will be reduced to zero by the unique  $x$  that satisfies

$$(8) \quad Rx = g$$

and the minimum value of the expression being minimized is just  $\alpha^2$ .

Summarizing these steps, we may write

$$\begin{aligned} (9) \quad \hat{S}^2 &= \min_x (Bx - y)^t H^{-1} (Bx - y) \\ &= \min_x \|L^{-1}Bx - L^{-1}y\|^2 \\ &= \min_x \|Rx - g\|^2 + \alpha^2 \\ &= \alpha^2 \end{aligned}$$

with the minimizing value of  $x$  being  $\hat{\xi}$  that satisfies

$$(10) \quad R\hat{\xi} = g.$$

Our unbiased estimate of  $\xi$  is  $\hat{\xi}$ . Our unbiased estimate of  $\phi^2$  is  $\hat{\phi}^2 = \hat{S}^2/(m-n)$ . Confidence intervals for  $\hat{\phi}^2$  are available by use of the  $\chi^2$  distribution as illustrated in Sec. 5.3.

The covariance matrix for  $\hat{\xi}$  is

$$(11) \quad \text{Cov}(\hat{\xi}) = \phi^2 V$$

where

$$(12) \quad V = (B^t H^{-1} B)^{-1} = (R^t R)^{-1} = R^{-1} R^{-t}$$

## 6.2. Combining sets of data

The total set of data defining the problem discussed in Sec. 6.1 consists of the vector  $y$ , and the matrices,  $B$  and  $H$ . Suppose two sets of data,  $(y_1, B_1, H_1)$  and  $(y_2, B_2, H_2)$ , have been acquired that are both assumed to derive from the same underlying parameter vector,  $\xi$ . Thus, for  $i = 1, 2$ , it is assumed that  $y_i$  is a noisy observation of  $B_i \xi$  with covariance matrix,  $\phi^2 H_i$ . Furthermore we assume the observations  $y_1$  and  $y_2$  are independent, in the sense that  $\text{Cov}(y_1, y_2) = 0$ .

Suppose the processing described in Sec. 6.1 has been applied to each of these data sets, obtaining estimates,  $\hat{\xi}_1$  and  $\hat{\xi}_2$ , as well as all of the intermediate and auxiliary quantities defined in Sec. 6.1.

We wish to consider the question of selecting from among these intermediate and auxiliary quantities those that will be useful in the computation of an estimate of  $\xi$  based on the combined data sets, and also to specify how these selected quantities are to be used to obtain such an estimate.

Supplementary quantities deriving from data sets 1 or 2 will be indicated by the symbol used in Sec. 6.1 subscripted with 1 or 2, respectively. Quantities based on combined data will be indicated by the symbol of Sec. 6.1 with no subscript. Thus we may begin by defining

$$(1) \quad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}$$

$$(2) \quad y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

$$(3) \quad H = \begin{bmatrix} H_1 & 0 \\ 0 & H_2 \end{bmatrix}$$

The combined problem can be characterized as that of minimizing the quantity  $S^2$  defined by

$$\begin{aligned}
 (4) \quad S^2 &= (Bx-y)^t H^{-1} (Bx-y) \\
 &= (B_1 x - y_1)^t H_1^{-1} (B_1 x - y_1) + (B_2 x - y_2)^t H_2^{-1} (B_2 x - y_2) \\
 &= \left\| \begin{bmatrix} R_1 \\ 0 \end{bmatrix} x - \begin{bmatrix} g_1 \\ \alpha_1 \end{bmatrix} \right\|^2 + \left\| \begin{bmatrix} R_2 \\ 0 \end{bmatrix} x - \begin{bmatrix} g_2 \\ \alpha_2 \end{bmatrix} \right\|^2 \\
 &= \| \tilde{R}x - \tilde{g} \|^2
 \end{aligned}$$

where

$$(5) \quad \tilde{R} = \begin{bmatrix} R_1 \\ 0 \\ R_2 \\ 0 \end{bmatrix}, \quad \tilde{g} = \begin{bmatrix} g_1 \\ \alpha_1 \\ g_2 \\ \alpha_2 \end{bmatrix}$$

The problem of minimizing  $S^2$  is now seen to be a linear least squares problem involving the matrix  $\tilde{R}$  and the vector  $\tilde{g}$ . A reasonable approach to solving this problem would be via the "QR" decomposition as in steps 6.1.(3) - (8). This "QR" decomposition can be written as

$$(6) \quad [\tilde{R} : \tilde{g}] = Q\tilde{U} = Q \begin{bmatrix} \hat{R} & \hat{g} \\ 0 & \hat{\alpha} \end{bmatrix}$$

where  $Q$  is a  $(2n+2) \times (2n+2)$  orthogonal matrix,  $\hat{R}$  is an  $n \times n$  nonsingular upper triangular matrix,  $\hat{g}$  is an  $n$ -vector,  $0$  denotes an  $n$ -dimensional row vector of zeros, and  $\hat{\alpha}$  is a scalar. We then have

$$(7) \quad S^2 = \| \tilde{R}x - \hat{g} \|^2 + \hat{\alpha}^2$$

so the the solution vector  $\hat{\xi}$  is given as the solution of

$$(8) \quad \hat{R}\hat{\xi} = \hat{g}$$

and the minimum value of  $S^2$  is

$$(9) \quad S^2 = \hat{\alpha}^2$$

The covariance matrix of  $\hat{\xi}$  is

$$(10) \quad \text{Cov}(\hat{\xi}) = \phi^2 \hat{V}$$

where

$$(11) \quad \hat{V} = (\hat{R}^t \hat{R})^{-1} = \hat{R}^{-1} \hat{R}^{-t}$$

Summarizing the above process we see that the upper triangular matrix,  $U$ , of Eq.6.1.(5) can be chosen as the quantity to save for each data set for use in later combining the data sets. The process of

combining data sets then amounts to stacking the saved U-matrices vertically, as in Eq.(5), and then computing the "QR" decomposition of this augmented matrix, as in Eq.(6), to obtain the U-matrix for the combined problem. The combined covariance matrix can be computed from the R-submatrix of the U-matrix, as in Eq.(11).

#### 6.2.1 Additional remarks on combining data sets

An alternative way to approach the minimization of the last expression in Eq. 6.2.(4) is to note that the minimizing vector  $\hat{\xi}$  is the unique solution of the normal equations

$$(1) \quad \hat{R}^t \hat{R} \hat{\xi} = \hat{R}^t \bar{g}$$

Thus the covariance matrix for  $\hat{\xi}/\phi$  can be expressed as

$$(2) \quad \hat{V} = (\hat{R}^t \hat{R})^{-1} = (R_1^t R_1 + R_2^t R_2)^{-1} = (V_1^{-1} + V_2^{-1})^{-1}$$

and  $\hat{\xi}$  can be expressed as

$$(3) \quad \begin{aligned} \hat{\xi} &= (\hat{R}^t \hat{R})^{-1} \hat{R}^t \bar{g} \\ &= \hat{V} (R_1^t \bar{g}_1 + R_2^t \bar{g}_2) \\ &= \hat{V} (R_1^t R_1 \hat{\xi}_1 + R_2^t R_2 \hat{\xi}_2) \\ &= (V_1^{-1} + V_2^{-1})^{-1} (V_1^{-1} \hat{\xi}_1 + V_2^{-1} \hat{\xi}_2) \end{aligned}$$

It is at least of mathematically aesthetic interest to note that Eqs.(2) and (3) can be written as

$$(4) \quad \hat{V}^{-1} = V_1^{-1} + V_2^{-1}$$

and

$$(5) \quad \hat{V}^{-1} \hat{\xi} = V_1^{-1} \hat{\xi}_1 + V_2^{-1} \hat{\xi}_2$$

Eqs.(2) and (3) show that as an alternative to the approach suggested in Sec. 6.2 of using the U-matrix as the object to be saved and updated as data sets are combined, one could instead use the  $\xi$ -vector and the V-matrix. These could then be updated as data sets are combined using Eqs.(2) and (3).

The use of U gives much better numerical stability and reliability than the use of V and  $\xi$ .

We assumed at the beginning of Sec. 6.1 that B was of rank n. In computational practice it is essential to recognize that, even though the matrix B may be of rank n, if its column vectors are nearly linearly dependent, i.e., if it has a large condition number (see Sec. 4.1), its performance in a computational process may be more like a matrix of lower rank. For this reason it is important to consider what will happen to a computational procedure if B is ill conditioned or has rank less than n.

The computation of the QR decomposition of a matrix is well defined and numerically stable regardless of the rank or condition number of the matrix being factored. Thus using the U-matrix as the fundamental object in processing separate data sets and combining them is a numerically stable process. If the R-submatrix of the U-matrix at some stage is ill-conditioned or rank-deficient then the quantities  $\hat{\xi}$  and V computed from the U-matrix at that stage are likely to be poorly determined or undeterminable. This only affects the  $\hat{\xi}$  and V at this stage, however. If this U-matrix is later combined with a U-matrix derived from another set of observations, it is possible that the R-submatrix of the new U-matrix will be better conditioned and reasonably accurate values of  $\hat{\xi}$  and V can be determined.

The U-matrix computed from some particular set of data will be essentially the same regardless of whether it is computed from all the data at once or through the combining of U-matrices previously computed for disjoint subsets of the data.

In contrast to the stability of the U-matrix approach, the approach based on Eqs.(2) and (3) is highly dependent on the order in which data is grouped. Error in the V-matrix at any stage will propagate to all following stages. The method fails completely if the V-matrix cannot be determined at some stage due to rank-deficiency of the underlying B-matrix, even though after more data is accumulated one might have a full-rank, and even well-conditioned, underlying B-matrix.

Updating methods based on Eqs.(2) and (3) were given in the early papers on "filtering" or "Kalman filtering" in the late 50's and early 60's. Instability of the type described here was recognized as a problem in those days. The QR decomposition was not widely known and understood in the early 60's, but by the late 60's it was being used in many algorithms of linear algebra. Its value in "filtering" was increasingly appreciated in the early 70's. A systematic treatment of "filtering" emphasizing the use of the U-matrix is given in [Bierman].

## REFERENCES

- Abramowitz, M., and I. A. Stegun, Handbook of Mathematical Functions, National Bureau of Standards, Applied Mathematics Series, No. 55 (1964).
- Anderson, T. W., An Introduction to Multivariable Statistical Analysis, Second Edition, Wiley (1984).
- Bierman, G. J., Factorization Methods for Discrete Sequential Estimation, Academic Press (1977).
- Golub, G. H., and C. Van Loan, Matrix computations, Johns Hopkins University Press (1983).
- Hoel, P. G., Introduction to Mathematical Statistics, Wiley (1955).
- Lawson, C. L., and R. J. Hanson, Solving Least Squares Problems, Prentice-Hall (1974).
- Plackett, R. L., Principles of Regression Analysis, Oxford University Press (1960), JPL Library No. QA 276 P697.

## INDEX

Affine transformation 9  
 Chi-squared distribution 22  
 Cholesky factorization 11  
 Circular bivariate normal distribution 21  
 Condition number 11  
 Confidence ellipsoid 21  
 Confidence region 21  
 Conjugate 10  
 Cov(b1,b2) 7  
 Cov(x) 12  
 Covariance 7  
 Covariance matrix 12  
 Degrees of freedom 8  
 Det(A) 12  
 Determinant 12  
 Distribution 4  
 Distribution function 4  
 E(x) 5, 12  
 Eigensystem factorization 11  
 Eigenvalue 10, 11  
 Eigenvector 11  
 Euclidean norm 9  
 Expected value 5  
 F distribution 26  
 Frequency function 4, 12  
 Full-rank 9  
 Gaussian distribution 20  
 Generalized inner product 10  
 Ill-conditioned 11  
 Independently distributed 13  
 Inner product 9  
 Joint frequency function 7  
 Least squares 15  
 Left identity 15  
 Left inverse 14  
 Linear transformation 9  
 Matrix 9  
 Mean 5, 12, 20  
 Minimum variance 7, 14, 15  
 Moment 19  
 N(\*,\*) 20  
 N(m; $\eta$ ,H) 20  
 Nonnegative definite 10  
 Nonsingular 9  
 Norm 9  
 Normal distribution 20  
 Normal equations 16  
 Orthogonal 10  
 Orthonormal 10  
 Outer product 9  
 Positive definite 10  
 Positive semidefinite 10  
 Probability 20, 21



Affine transformation 10  
Probability density function 4, 12  
QR factorization 11  
Random variable 4  
Rank 9, 13  
Full 9  
Rank-deficient 9  
Singular value decomposition 11  
Singular values 11  
Spectral norm 10  
Standard deviation 5, 20  
Student's t 24  
Symmetric 10, 12  
Unbiased 7, 14, 15  
Var(x) 6  
Variance 5  
Vector 9  
Well-conditioned 11